Applying Heuristic Evaluation to Human-Robot Interaction Systems

Edward Clarkson and Ronald C. Arkin

College of Computing and GVU Center, Georgia Institute of Technology
801 Atlantic Drive, Atlanta, GA 30332-0280

{edcclark, arkin}@cc.gatech.edu

ABSTRACT

Though attention to evaluating human-robot interfaces has increased in recent years, there are relatively few reports of using evaluation tools during the development of human-robot interaction (HRI) systems to gauge and improve their designs—possibly due to a shortage of suitable evaluation techniques. Heuristic evaluation is a technique suitable for such applications that has become popular in the human-computer interaction (HCI) community. However, it requires usability heuristics applicable to the system environment. This work contributes a set of heuristics appropriate for use with HRI systems, derived from a variety of sources both in and out of the HRI field. Evaluators have successfully used the heuristics on an HRI system, demonstrating their effectiveness against standard measures of heuristic effectiveness.

Categories and Subject Descriptors

H.5.2 [User Interfaces]: Evaluation/methodology

Keywords

Heuristic evaluation, discount evaluation, usability testing, HRI.

1. INTRODUCTION

The attention paid to human-robot interaction (HRI) issues has grown dramatically as robotic systems have become more capable and as human contact with those systems has become more commonplace. Along with the development of robotic interfaces, there has been an increase in the evaluation of these systems. HRI researchers have employed a variety of evaluation styles in their work; they can evaluate their systems *summatively* (i.e., after-the-fact) or *formatively* (i.e., during system development). However, there have been relatively few accounts of formative applications or uses of *discount* (low-cost) techniques—two evaluation classes that have been explored extensively in traditional human-computer interaction (HCI) research. Discount methods used in formative evaluations can be powerful tools. Not only do they take small amounts of time or resources, but they can catch both major and minor problems early in the development cycle. Identifying problems earlier in system development drastically reduces the cost of fixing them in both commercial and research project settings.

One discount evaluation technique is heuristic evaluation (HE) [13] [14], a method that has become popular in both the professional and academic HCI communities. HE consists of a small group of evaluators who examine an interface using a set of heuristics as a guide for their inspection. Its low cost makes it well suited to formative evaluations. However, the application of HE to a problem depends on the availability of a set of heuristics that are applicable to the problem domain.

This work presents our work to synthesize such a set of heuristics that are specific to HRI systems. This allows for the successful application of HE to HRI systems and also encourages the use of formative evaluations in HRI system design. Our development procedure is based on accepted methodology from previous adaptations of heuristic evaluation (HE) to new problem domains [2] [12], and takes inspiration for the heuristics themselves from a variety of existing works in HRI [8] [22] [23] [24] and related fields [2] [12] [15]. We present our application of our derived heuristics to the evaluation of an example HRI system, which shows that 3-5 evaluators using the set find 40-60% of known usability problems (the standard test for heuristic effectiveness). We also examine the differences between evaluators specializing in HCI and robotics; we find that there is no statistically significant difference in the quantities or the severity of the problems found by the evaluator groups. The final result of our work yields a validated set of HRI heuristics, suitable for use by robotics researchers with little or no previous evaluation experience.

2. RELATED WORK

Until recently, evaluation on HRI systems has not received its due attention. Georgia Tech's Skills Impact Study notes that most researchers have contributed "lip service" to evaluation, but not many actual experimental studies [7]. Yanco et al. [24] make similar statements, noting that scant work has gone into assuring that HRI displays and interactions controls are at all intuitive for their users.

However, more researchers have recently come to recognize the need for evaluation and developing evaluation guidelines. One of the recommendations put forth as part of a case study on urban search and rescue (USAR) at the World Trade Center site [4] was for additional research in perceptual user interfaces. A recent DARPA/NSF report also proposes research in evaluation methodologies and metrics as one productive direction for future research [3], citing the need for evaluation methods and metrics that can be used to measure the development of human-robot teams.

Among evaluation research that has been conducted, a common approach has been case studies of various systems in the field [17] [20] [24]. The number of controlled lab studies of HRI interfaces has also increased over recent years, with most studies focused on comparing various interface alternatives for teleoperation [11] [18] or plan specification software [6]. Various subjective and objective measures have been used in these studies. Qualitative pre- and post-test subject questionnaires, interviews and

experimenter observation are common (much like many in-situ HCI studies). Quantitative, empirical measurements (in both field and lab studies) have included task completion time, error frequency, and other *ad hoc* gauges particular to the task environment [24]. Olsen and Goodrich have suggested a set of six interrelated metrics for judging different aspects of a human-robot interaction (e.g., attention demand, neglect tolerance) [19]. The NASA Task Load Index (NASA-TLX) measurement scale has also been used as a quantitative measure of operator mental workload during robot teleoperation and partially autonomous operation [11] [20].

2.1 Formative and Summative Evaluation

A common thread (with a few exceptions, e.g., [17]) among HRI evaluation literature is the focus on formal summative studies and techniques. Designers use summative evaluations to judge the outcome of a design implementation and formative evaluations to assess preliminary design products with the intention of guiding the design or implementation itself. Note that this distinction concerns how and when a technique is applied and not the method *per se*. However, many evaluation techniques are obviously better suited to one application or the other.

We have already noted that research is needed into techniques that can be used to judge the progress of HRI systems. Progress can occur on both large (the advancement of HRI as a field) and small (the development of an individual project) scales, but both interpretations are important to HRI. We can make two observations about the state-of-the-art in HRI:

- There are relatively few validated tools and techniques for evaluating HRI systems.
- There have been few reports of performing formative evaluations (or other principles of user-centered design) in the HRI literature [1].

Cleary, the former situation is a large contributing factor to the latter. Hence, there is a need for evaluation methods that are both suited to formative studies and have been successfully demonstrated specifically on HRI applications.

2.2 Discount and Heuristic Evaluation

Discount evaluation techniques are methods that are designed explicitly for low cost (in terms of manpower and time). Because of these properties, discount evaluations are often applied formatively. One such approach whose popularity has grown rapidly since its introduction in the HCI community is heuristic evaluation (HE). HE is a type of usability inspection method, which is a class of techniques involving evaluators examining an interface with the purpose of identifying usability problems. This class of methods has the advantage of being applicable to a wide range of prototypes, from detailed design specifications to fully functioning systems. HE was developed by Nielsen and Molich and has been empirically validated [13], [14]. In accordance with its discount label, it requires only a few (three to five) evaluators who are not necessarily HCI (or HRI) experts (though it is more effective with training).

The principle behind HE is that individual inspectors of a system do a relatively poor job, finding a fairly small percentage of the total number of known usability problems. However, Nielsen has shown evaluators have a wide variance in the problems they find, which means the results of a small group of evaluators can be aggregated with little duplication to uncover a large number of bugs. Briefly, the HE process in particular consists of the following steps [15]:

- The group that desires a heuristic evaluation (such as a design team) performs preparatory work, generally in the form of:
    1. Creation of problem report templates for use by the evaluators.
    2. Customization of heuristics to the specific interface being evaluated. Depending on what kind of information the design team is trying to gain, only certain heuristics may be relevant to that goal. In addition, since canonical heuristics are (intentionally) generalized, heuristic descriptions given to the evaluators can include references and examples taken from the system in question.
- Assemble a small group of evaluators (Nielsen recommends three to five) to perform the HE. These evaluators do not need any domain knowledge of usability or interface design.
- Each evaluator **independently** assesses the system in question and judges its compliance with a set of usability guidelines (the heuristics) provided for them.
- After the results of each assessment have been recorded, either the evaluators or the experimenter aggregate the overall results and assign severity ratings to the various usability issues.

HE has been shown to find 40 – 60% of usability problems with just three to five evaluators (hence Nielsen's recommendation), and a case study showed a cost to benefit ratio of 1:48 (as cited in [15]). For those reasons, among others, HE has proven to be very popular in both industry and research. Of course, the results of an HE are highly subjective and probably not repeatable. However, it should be emphasized that this is not a goal for HE; its purpose is to provide a proven evaluation framework that is easy to teach, learn and perform while also uncovering a significant percentage of usability problem. Its value thus comes from its low cost, its applicability early in the design process, and the fact that even those with little usability experience can administer a useful HE study. These features, which make it such a popular HCI method, also make it useful for gauging HRI systems.

3. HEURISTIC EVALUATION FOR HRI

The problem with applying heuristic evaluation to HRI, however, is the validity of using existing heuristics for HRI. Scholtz states that HRI is "fundamentally different" from normal HCI in several aspects, and Yanco et al. acknowledge HE as a useful HCI method, but rejects its applicability to HRI because Nielsen's heuristics are not appropriate to the domain. There are many issues listed as differentiating factors between HRI and HCI/HMI, including complex control systems, the existence of autonomy and cognition, dynamic operating environments, varied interaction roles, multi-agent and multi-operator schemes, and the embodied nature of HRI systems. Despite these domain variations, the obvious question is whether it is possible to form a new set of heuristics that are pertinent to HRI systems?

If we can rely on the HCI literature, the answer is 'yes'. Alternative sets of heuristics have already been developed for

domains outside the traditional on-the-desktop software realm. Confronted with similar problems—Nielsen's heuristics did not address the focus of computer-supported cooperative work (CSCW) applications (teamwork)—Baker et al. adapted heuristics for use with the evaluation of groupware applications [2]. Similarly, Mankoff and her collaborators produced a heuristic list for ambient displays [12].[1] In each case, the new heuristic lists were developed and validated according using similar methodologies, each based in part on Nielsen's original work.

The development of the CSCW guidelines was a lengthy process, begun in 1999 [8] and continued until 2002. Initially, the CSCW heuristics were based on the Locales framework of social interaction. Modification to the list continued via case studies and the use of the mechanics of collaboration framework. The authors also undertook an empirical study involving a large number (27) novice and expert inspectors using their heuristics to evaluate two groupware applications. Their results were similar to that of Nielsen's: an average group of three to five inspectors from their overall evaluators found between 40% and 60% of known usability problems, the consensus benchmark for an effective heuristic set.

Mankoff [12] pursued a slightly different approach, but still in keeping with previous work. The ambient guidelines were developed more rapidly, and based directly on the standard list from Nielsen. The list underwent a short cycle of modification based on informal surveys and pilot studies. Sixteen subjects were then recruited to perform an HE of two peripheral displays. Eight inspectors used the ambient heuristics and eight Nielsen's in a between-subjects comparison, which showed that the ambient heuristics found more and more severe problems than the standard set. After the study it was noted that Nielsen's heuristics did find some problems never identified in the ambient set. To remedy that, the authors repeatedly selected the heuristic (chosen from among both sets) that accounted for the largest number of severe problems until none were left. This process formed their final list of ambient heuristics, which included nearly all of the ambient set and half of Nielsen's.

The problem domains for each of these adaptations are also noteworthy. CSCW applications are entirely concerned with facilitating teamwork, organizing group behavior and knowledge, and providing support for simultaneous interaction—all issues that are similar to those that separate HCI from HRI in multi-agent and multi-operator settings. Likewise, ambient devices strive to convey information without interrupting the users attention, and do so through a variety of software or hardware form factors; evaluations of HRI systems face similar variability in trying to judge how well a system maintains operator awareness of sensor data or determine the effects of a robot's physical appearance

3.1 HRI Heuristic Development

Following the methodology similar to Baker et al. and Mankoff et al., (which were based in turn on Nielsen's method for creating his initial list), our process for heuristic development consists of three broad steps:

- Create an initial list of HRI heuristics via brainstorming and synthesizing existing lists of potentially applicable heuristics.
- Modify the initial list based on pilot studies, consultation with other domain experts, and other informal techniques.
- Validate the modified list against existing HRI systems.

There are a number of bases from which to develop for potential HRI heuristics: Nielsen's canonical list [15], HRI guidelines suggested by Scholtz [22], and elements of the ambient [12] and CSCW heuristics as well [2]. Scholtz's issues in particular are almost directly applicable as heuristics, although they do not seem to be proposed for that purpose. They have been used as "high-level evaluation criteria," however, in a manner that bears some similarity to heuristic evaluation. Sheridan's challenges for the human-robot communication [23] can also be considered issues to be satisfied by an HRI system.

These lists and the overall body of work in HRI provide the basis for heuristics applicable to both multi-operator and multi-agent systems; however, for this work we have limited our focus to single operator, single agent settings. There are several reasons for doing so: it narrows the problem focus; the complexity introduced by multi-operator or -agent systems is to a large degree orthogonal to the base case; the validation of single human/robot systems is a wise first step since lessons learned during this work can be applied to further development.

Our initial list is based on the distinctive characteristics of HRI, and should ideally be pertinent to systems ranging from normal windowed software applications to less traditional interfaces such as that of the Sony entertainment robots or the iRobot Roomba vacuum cleaner. Our list should also apply equally well to purely teleoperated machines, monitored autonomous systems, and everything in between. This is a feasible goal, as in some ways, what makes a robotic interface (no matter what the form) effective is no different than what makes anything else usable, be it a door handle or a piece of software.

To accomplish these goals, we can identify issues relevant to the user that are common to all of these situations. Norman emphasizes a device's ability to communicate its state to the user as an important characteristic [16]. Applied to a robot, an interface then should make evident various aspects of the robots status—what is its pose? What is its current task or goal? What does it know about its environment, and what does it not know? Parallel to the issue of *what* information should be communicated is *how* it is communicated. The potential complexity of robotic sensor data or behavior parameters is such that careful attention is due to what exactly the user needs out of that data, and designing an interface to communicate that data in the most useful format.

Many of these questions have been considered as a part of the heuristic sets mentioned previously, and we leverage that experience by taking elements in whole and part from those lists to form our own attempt at an HRI heuristic set. The inspirational source or sources before adaptation accompany each heuristic in Table 1. Heuristics 1, 2, and 3 all deal with the handling of information in an HRI interface, representative of the importance of data processing in HRI tasks. Eight signifies the potential importance of emotional responses to robotic systems. Number five is indicative of interfaces ability to immerse the user in the system, making operation easier and more intuitive. Four, five

[1] *Ambient displays* are "…aesthetically pleasing displays of information which sit on the periphery of a user's attention" [12].

1. **Sufficient information design** *(Scholtz, Nielsen)*

The interface should be designed to convey "just enough" information: enough so that the human can determine if intervention is needed, and not so much that it causes overload.

2. **Visibility of system status** *(Nielsen)*

The system should always keep users informed about what is going on, through appropriate feedback within reasonable time. The system should convey its world model to the user so that the user has a full understanding of the world as it appears to the system.

3. **Appropriate information presentation** *(Scholtz)*

The interface should present sensor information that is clear, easily understood, and in the form most useful to the user. The system should utilize the principle of recognition over recall.

4. **Match between system and the real world** *(Nielsen, Scholtz)*

The language of the interaction between the user and the system should be in terms of words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

5. **Synthesis of system and interface** *(None)*

The interface and system should blend together so that the interface is an extension of the system itself. The interface should facilitate efficient and effective communication between system and user and vice versa.

6. **Help users recognize, diagnose, and recover from errors** *(Nielsen, Scholtz)*

System malfunctions should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution. The system should present enough information about the task environment so that the user can determine if some aspect of the world has contributed to the problem.

7. **Flexibility of interaction architecture** *(Scholtz)*

If the system will be used over a lengthy period of time, the interface should support the evolution of system capabilities, such as sensor and actuator capacity, behavior changes and physical alteration.

8. **Aesthetic and minimalist design** *(Nielsen, Mankoff)*

The system should not contain information that is irrelevant or rarely needed. The physical embodiment of the system should be pleasing in its intended setting.

Table 1 – Initial HRI heuristics.

and six all deal with the form communication takes between the user and system and vice versa. Finally, Heuristic 7 reflects the longevity and adaptability often required of HRI platforms .

Since the heuristics are intended for HRI systems, they focus only on the characteristics distinct to HRI. Many human-robot interfaces (especially those that are for the most part traditional desktop software applications) can and do have usability problems that are associated with 'normal' HCI issues (e.g., widget size or placement), but these problems can be addressed by traditional HCI evaluations.

3.2 HRI Heuristic Validation

Our validation plan is similar to that described in both Baker and Mankoff:

- Create an initial list of HRI heuristics via brainstorming and synthesizing existing lists of potentially applicable heuristics (accomplished above).
- Use the heuristics in an evaluation of an HRI system.
- Hypothesize that a small number of evaluators using the heuristics will uncover a large percentage of known usability problems.
- Modify the initial heuristic list based on the results.

It is necessary to have a relatively large group of evaluators for the purposes of assessing the heuristics. Though HE generally requires only a few (3-5) evaluators, a larger group enables us to test whether an arbitrary subset of the overall group can indeed uncover a significant percentage of usability problems.

3.2.1 Experimental Procedure

We lacked an active project using an HRI system appropriate for such an evaluation, so we created an *ad hoc* system and problem for this work. We chose the RoboCup Rescue[2], an annually-held worldwide USAR competition, as our problem environment. The contest is held in an indoor arena designed to mimic a portion of an urban area after a large-scale disaster such as an earthquake. We chose a robot based on the Segway RMP platform as the HRI system to be evaluated (see figure Figure 1). The system as presented to users was teleoperated using the Mobile Robot Lab's MissionLab software package and a standard PC analog joystick controller. It contained two major sensory systems, a pair of SICK laser rangefinders (mounted parallel to the floor) and a forward-mounted optical camera.

We have presented only an outline of the system in question; the contribution of this work is not the system or its evaluation but the results of that evaluation as it informs the development of our heuristics. Indeed, our HRI system and problem environment are not particularly well-suited for each other by design. The purpose of a HE is to uncover interaction problems (the more severe the better), and a problem/system mismatch ensures their presence for evaluators. We report the specificities of the problems indicated by our evaluation only insofar as they inform the development and validation of our heuristics.

We recruited ten HCI and robotics graduate students to serve as our evaluator team. Two did not complete the entirety of the evaluation and are ignored henceforth. The eight remaining had a mean age of 28 years and five of them were female. Three evaluators had a specialization in robotics and the other five specialized in HCI.

The evaluation consisted of preparing and distributing a packet of written information about the evaluation. This included an introduction and summary of both the HRI system and the problem environment, and collection of problem report templates. The problem report templates provided pre-labeled fields for a problem summary or title, a detailed description and an indication of which heuristic the problem violated. We also discussed the information contained in the packet in a meeting with all the evaluators. This included an introduction to the general heuristic evaluation procedure, a presentation on the robot's sensors and capabilities, the RoboCup rescue competition and rules, and a live demonstration of the operation of the system.

[2] http://www.rescuesystem.org/robocuprescue/

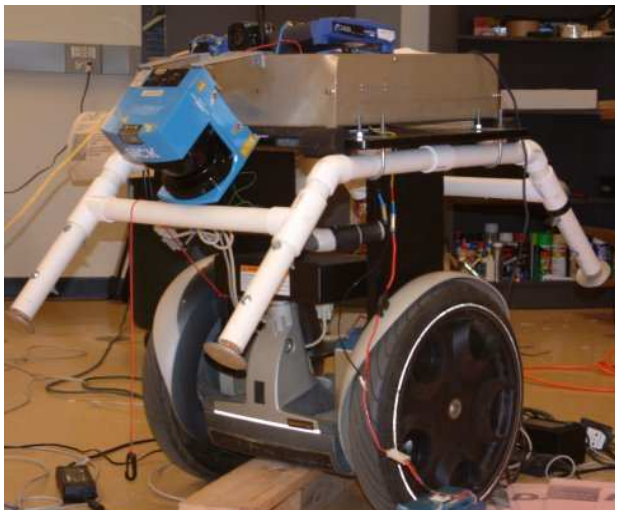

Figure 1 – The robot portion of the HRI system.

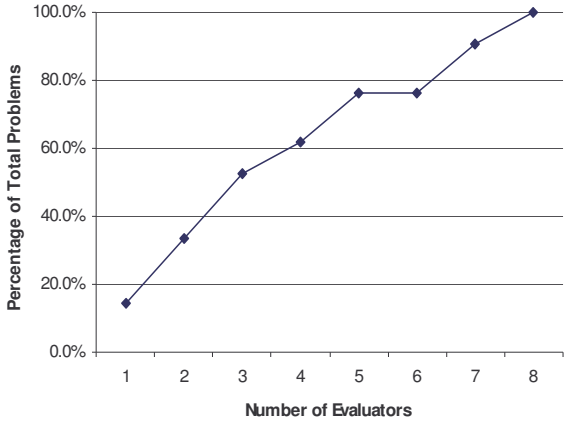

Figure 2 – Percentage of known problems found with increasing number of evaluators.

Evaluators were encouraged to return within a week as many problem reports as they deemed appropriate. We also instructed them to prepare their problem reports independently. We did not suggest specific time-on-task guidelines for completing the problem reports.

4. RESULTS

The evaluators as a group returned 59 problem reports. Individual counts ranged between 5 and 10. We synthesized the results by combining duplicate problem reports. Such duplicates are sometimes obvious ("Only one camera; doesn't move; doesn't cover 360 deg." and "Camera direction/control [is] fixed position, can't move it.") Other duplications are more subtle: "Map doesn't show orienting features, can't mark locations of interest. No ability to save history of movements" and "Need indication of how many victims found and where, hazards and locations, running point total" reflect different aspects of the same problem. That is, the system does not effectively provide historical data about the significant environmental features.

After synthesizing the results in this manner, we identified 21 unique problems and assigned severity ratings to each of them using a standard rating system of 0-4 with 4 being the most severe and 0 being a non-problem. Evaluators found 11 severe problems (ratings or 3 or 4) and 10 minor problems (ratings of 1-2). Average severity across all 21 problems was 2.52. There were no non-problems reported. The average single evaluator found 29% of the known problems, a figure comparable to those reported by Nielsen [15] and Baker [2].

Table 2 shows a representation of how problem identification is distributed across the different evaluators. Evaluators are represented by columns and ordered from least to most successful (measured by the number of unique problems reported). Each row signifies a unique problem, and they are ordered according to severity. The chart shows that there is substantial spread among the different evaluators and that even relatively unsuccessful evaluators are able to identify severe problems. Furthermore, unsuccessful evaluators in some cases were the *only* inspector to identify a particular problem.

Similarly, Figure 2 charts the increasing percentage of known problems found with additional evaluators (with additions to the total from the least to most successful evaluator). Most notably, it shows that the heuristics passed the canonical HE test: 3-5 evaluators identify at least 40-60% of the known problems. An inspection of other randomly-ordered graphs showed similar

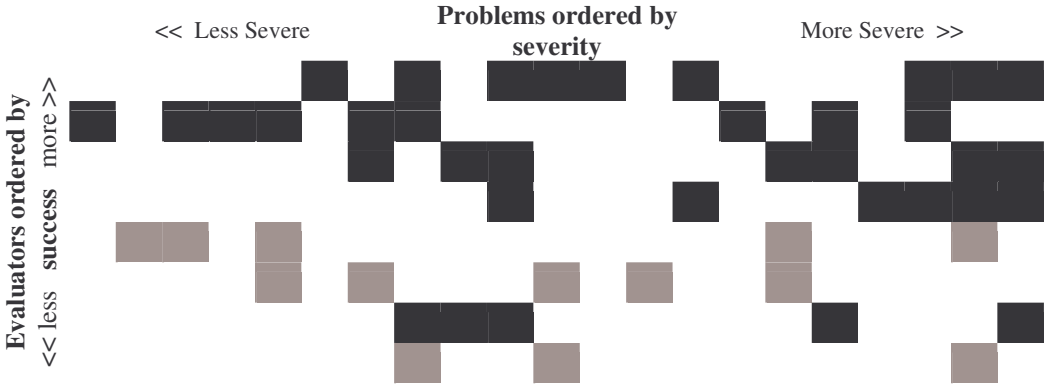

Table 2 – A chart of the problems found by each evaluator. A filled square indicates the problem corresponding to that row was identified by the evaluator corresponding to that column. Black rows are HCI specialists; gray rows are roboticists.

results.

Many projects in robotics have limited access to HCI specialists, or at least have much easier access to roboticists: for example, student teams at competitions like RoboCup Rescue. As a result, we were also interested if there are any differences between evaluators with a background in robotics ("group R") and HCI ("group H"). In our evaluation, group H found 7.6 unique problems against a mean of 5.3 for group R. The average severity of the problems found by each group was almost identical at 2.53 for group H and 2.51 for group R ($p = 0.90$). A two-tailed t-test indicates the difference in problem totals is marginally significant ($p = 0.06$). Notable is the fact that all of group H had participated in and 80% had themselves conducted an HE prior to our study; only one member of group R had participated and conducted in an HE. As such, familiarity with the HE process may be a contributing factor to this result (in addition to a relatively small sample size). However, even with apparently less effective evaluators, the three roboticists identified 43% of the total problems, still within the standard for an acceptable HE process. This indicates that teams of "regular" roboticists can perform effective HEs with little or no prior experience, a significant advantage for real-world projects.

4.1 Discussion

A number of issues with our heuristics arose explicitly via evaluator comments or implicitly through their problem responses. One of the most severe problems with our example HRI system is that its sensor capabilities simply are not adequate to perform the tasks expected in the competition. However, none of our heuristics plainly mention checking system capabilities against expected tasks (though heuristic 7 comes close). Likewise, many of the other most severe problems with our HRI system relate to the difficulty in maintaining an accurate mental model of the robot and its surroundings. This is sometimes termed 'situational awareness' and has been identified as an important aspect of HRI systems [5]. Heuristic 2 touches on this idea, but does not use the term 'situational awareness' explicitly.

Two other matters may be of interest to other researchers working with the evolution of HRI heuristics. First, evaluator incentives are an important consideration. Our experience in many ways was similar to that of Baker's, whose team found that, surprisingly, novice evaluators outperformed groupware experts. However, after reexamining their procedures they noted that their novice group were students who were assigned the HE as part of a class assignment; in contrast, their experts were volunteer participants motivated only by a willingness to help. Our evaluators were similarly altruistic, and another reason that group H was more productive may have been from an expectation of *quid pro quo* volunteerism—HCI specialists tend to perform more evaluations and thus may put forth more effort in hopes of other returning the favor in the future.

We also found our use of an *ad hoc* system for evaluation purposes to be limiting in some ways. Since we did not employ the system for its purported use (i.e., compete in RoboCup Rescue), we cannot have a true appreciation for the full scope of the problems and issues that comes with actual familiarity. Similarly, because there were so many obvious mismatches between the task and our HRI system, it is difficult to gauge whether the existing problems could have been qualitatively different from ones in a more realistic scenario.

Goodrich and Olsen have also proposed seven principles for effective HRI systems [8] based on their metrics for measuring HRI system performance [19]. They are: implicitly switch interfaces and autonomy modes; let the robot use natural human cues; manipulate the world instead of the robot; manipulate the relationship between the robot and world; let people manipulate presented information; externalize memory; and help people manage attention. Many of these principles are covered explicitly or implicitly in our initial heuristic set though they were not used in their original development. For example, "use natural cues" is another way of saying "the language of the interaction between the user and the system should be in terms of words, phrases and concepts familiar to the user" (heuristic 4). To "directly manipulate the world" requires an interface which acts simply as an extension of the HRI system (heuristic 5).

4.2 Updated HRI Heuristics

Though our heuristics performed well in our tests, our findings mentioned in our discussion above led us to revise our heuristics, clarifying them by rewording or adding various passages. The final results are presented in

Table 3. We have added an overt mention of situational awareness to heuristic 2; added language to heuristic 3 and 5 to reflect better several of Goodrich and Olsen's principles; re-titled heuristic 4 with their "use natural cues" phrase, which is clearer and more succinct than Nielsen's original heading; and added language to heuristic 7 to ensure a check for appropriate hardware capabilities.

5. CONCLUSIONS

We have noted that there has been little mention of formative evaluation in HRI research, due in part to a lack of tested methods for conducting them on HRI systems. The utility of formative evaluations is strong motivation for the use of such methods in HRI. Heuristic evaluation, a usability inspection method from HCI, is ideal for formative applications. Previous work has validated the concept of adapting HE to new problem domains, and those problem domains share some of the differences between traditional HCI and HRI system—indicating that adapting a set of heuristics for HRI is a fruitful endeavor. To that end, we have proposed an initial set of heuristics intended for single operator, single agent human-robot interaction systems, validated them against an example HRI system and amended the set based on our experience with the evaluation. Our tests also indicate no significant differences between robotics and HCI evaluators, indicating teams of roboticists can independently perform successful HEs.

Future work in this area is promising. Certainly, additional use of these heuristics will improve both the targeted systems and the heuristics themselves. Their indirect promotion of formative evaluation can improve the efficiency and efficacy of HRI development efforts. Their continued use may also inform the development of heuristics for multi-robot or -human settings. Other more specialized contexts which may benefit from their own heuristics might include Scholtz's interaction roles [21] or affective/sociable robots.

1. **Sufficient information design**

   The interface should be designed to convey "just enough" information: enough so that the human can determine if intervention is needed, and not so much that it causes overload.

2. **Visibility of system status**

   The system should always keep users informed about what is going on, through appropriate feedback within reasonable time. The system should convey its world model to the user so that the user has a full understanding of the world as it appears to the system. The system should support the user's situational awareness.

3. **Appropriate information presentation**

   The interface should present sensor information that is clear, easily understood, and in the form most useful to the user. The system should utilize the principle of recognition over recall, externalizing memory. The system should support attention management.

4. **Use natural cues**

   The language of the interaction between the user and the system should be in terms of words, phrases and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

5. **Synthesis of system and interface**

   The interface and system should blend together so that the interface is an extension of the system, the user and by proxy, the world. The interface should facilitate efficient and effective communication between system and user and vice versa, switching modes automatically when necessary.

6. **Help users recognize, diagnose, and recover from errors**

   System malfunctions should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution. The system should present enough information about the task environment so that the user can determine if some aspect of the world has contributed to the problem.

7. **Flexibility of interaction architecture**

   If the system will be used over a lengthy period of time, the interface should support the evolution of system capabilities, such as sensor and actuator capacity, behavior changes and physical alteration. Sensor and actuator capabilities should be adequate for the system's expected tasks and environment.

8. **Aesthetic and minimalist design**

   The system should not contain information that is irrelevant or rarely needed. The physical embodiment of the system should be pleasing in its intended setting.

Table 3 – Revised HRI heuristics.

6. ACKNOWLEDGMENTS

Many thanks are due to our volunteer evaluators for their generous contributions of time and effort. Thanks also to Yoichiro Endo for his review of an early draft of this work.

7. REFERENCES

[1] Adams, J. Critical Considerations for Human-Robot Interface Development. In *Proceedings of the 2002 AAAI Fall Symposium on Human-Robot Interaction*, pp. 1-8.

[2] Baker, K., Greenberg, S. and Gutwin, C. Empirical development of a heuristic evaluation methodology for shared workspace groupware. In *Proceedings of CSCW '02*, pp. 96–105.

[3] Burke, J., Murphy, R.R., Rogers, E., Scholtz, J., and Lumelsky, V. Final Report for the DARPA/NSF Interdisciplinary Study on Human-Robot Interaction. *IEEE Systems, Man and Cybernetics* Part C (34) 2, pp. 103-112.

[4] Casper, J., and Murphy, R. Human-Robot Interactions during the Robot-Assisted Urban Search and Rescue Response at the World Trade Center. *IEEE Transactions on Systems, Man and Cybernetics* Part B, (33) 3, pp. 367-385.

[5] Drury, J., Scholtz, J. and Yanco. H. Awareness in Human-Robot Interactions. In *Proceedings of IEEE Conference on Systems, Man and Cybernetics '03*, pp. 912-918.

[6] Endo, Y., MacKenzie, D.C., and Arkin, R. Usability Evaluation of High-Level User Assistance for Robot Mission Specification. *IEEE Transactions on Systems, Man, and Cybernetics* Part C, (34) 2, pp. 168-180.

[7] Georgia Tech College of Computing and Georgia Tech Research Institute. Real-time Cooperative Behavior for Tactical Mobile Robot Teams; Skills Impact Study for Tactical Mobile Robot Operational Units. DARPA report, 2000. Available at http://www.cc.gatech.edu/ai/robot-lab/tmr/skillsassessment.pdf

[8] Goodrich, M. and Olsen, D. Seven Principles of Efficient Human Robot Interaction. In *Proceedings of PERMIS '03.*

[9] Greenberg, S., Fitzpatrick, G., Gutwin, C. & Kaplan, S. Adapting the Locales framework for heuristic evaluation of groupware. In *Proceedings of OZCHI '99*, pp. 28-30.

[10] Haigh, K. and Yanco, H. Automation as Caregiver: A Survey of Issues and Technologies. In *Proceedings of AAAI '02 Workshop on Automation as Caregiver: The Role of Intelligent Technology in Elder Care*, pp. 39-53.

[11] Johnson, C., Adams, J. and Kawamura, K. Evaluation of an Enhanced Human-Robot Interface. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics '03*, pp. 900-905.

[12] Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Ames, M., Lederer, S. Heuristic evaluation of ambient displays. In *Proceedings of CHI '03*, pp. 169-176.

[13] Molich, R., and Nielsen, J. Improving a human-computer dialogue. *Communications of the ACM* (33) 3 (March), pp. 338-348.

[14] Nielsen, J. Enhancing the explanatory power of usability heuristics. In *Proceedings of CHI '94*, pp. 152-158.

[15] Nielsen, J. "How to Conduct a Heuristic Evaluation." http://www.useit.com/papers/heuristic/heuristic_evaluation.html.

[16] Norman, D. *The Design of Everyday Things*. Doubleday, New York, 1990.

[17] Nourbakhsh, I., Bobenage, J., Grange, S., Lutz, R., Meyer, R. and Soto, A. An affective mobile robot educator with a full-time job, *Artificial Intelligence* 114 (1–2), pp. 95–124.

[18] Olivares, R., C. Zhou, J. Adams, and B. Bodenheimer. Interface Evaluation for Mobile Robot Teleoperation. In *Proceedings of the ACM Southeast Conference '03*, pp. 112-118.

[19] Olsen, D. and Goodrich, M. Metrics for Evaluating Human-Robot Interactions. In *Proceedings of PERMIS '03.*

[20] Schipani, S. An Evaluation of Operator Workload, During Partially-Autonomous Vehicle Operation. In *Proceedings of PERMIS '03.*

[21] Scholtz, J. Theory and Evaluation of Human-Robot Interaction. In *Proceedings of HICSS '03*.

[22] Scholtz, J. Evaluation methods for human-system performance of intelligent systems. In *Proceedings of PERMIS '02*.

[23] Sheridan, T. Eight ultimate challenges of human-robot communication. In *Proceedings of RO-MAN '97*, pp. 9–14.

[24] Yanco, H., Drury, J. and Scholtz, J. Beyond Usability Evaluation: Analysis of Human-Robot Interaction at a Major Robotics Competition. *Journal of Human-Computer Interaction*, (19) 1 and 2, pp. 117-149.